Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach

Chitra Arjun, Mr.Anto S

***Abstract*— In recent years, support vector machines (SVMs) have shown good performance in a number of application areas. The existing system is concentrated on the discovery of risk of having pre-diabetes or undiagnosed diabetes and to facilitate people decide whether they should see a physician for further evaluation. It is also focused on both the noninvasive and metabolic factors, which should require blood sampling and laboratory measurements, such as high density lipoprotein (HDL), and cholesterol (CHOL). However the existing system ahs issue with prediction results by using c4.5, naïve bayes tree and neural network algorithms. To avoid the above mentioned issue we go for proposed system. In proposed scenario, we introduced an efficient algorithm named as Support Vector Machine (SVM) which is utilized to screen diabetes, and an ensemble learning module is added. It turns the "black box" of SVM decisions into comprehensible and transparent rules, and it is also useful for solving difference problem. The proposed system is used to develop an ensemble system for diabetes diagnosis. Specifically, the rules are extracted from the SVM algorithm and it is applied to provide comprehensibility and transparent representation. These rule sets can be regarded as a second opinion for diagnosis and a tool to screen the individuals with undiagnosed diabetes by lay users. From the experimental result, we can conclude that the proposed system is better than the existing scenario in terms of reduction of the incidence of diabetes and its complications.**

***Index Terms*— diagnosis of diabetes, ensemble learning, random forest (RF), rule extraction, support vector machines (SVMs)**

I. INTRODUCTION

Diabetes is a disease in which the body does not generate or correctly use insulin, the hormone that ‘‘unlocks" the cells of the body, allowing glucose to go into and fuel. Diabetes increases the risks of initial kidney disease, loss of sight, nerve injure, blood vessel damage and it contributes to heart disease. The cause of diabetes continues to be a ambiguity, although both genetics and ecological factors such as obesity and be short of of exercise come out to take part in roles.

Some of the most accepted classification techniques are based on the formation of propositional if-then rules from prelabeled training data. These methods are in principle that can provide an entirely transparent classification decision, but, in fact, their performance and comprehensibility frequently bear in cases of high-dimensional data and continuously valued attributes. Another trendy family of classifiers exemplified by support vector machines (SVMs) and artificial neural networks (ANNs) builds a mathematical form of the data that often performs much better in these situations. However, these methods construct black box models with little or no explanation capacity. In application areas such as medical analysis, there is a obvious need for an description component to be coupled with classification decisions in order to aid the approval of these methods by users[1].

One may disagree that in spite of all the hard work in the grassland of rule extraction from ANNs, there is no clear proof that this area of study was victorious. This dispute is valid to a great extent and can be mostly attributed to the refuse in the use of ANNs in the late 1990s, as they were largely outdated by SVMs because of their superior performance in a number of ordinary applications. Another motivation is that the popular of rule extraction algorithms from ANNs were narrow to a specific network type or architecture . However, in the near future at least, we can estimate an augment in the development and use of SVM rule extraction techniques corresponding with the developments and use of SVMs in a diversity of applications. Furthermore, a number of capable SVM rule extraction algorithms published to date are both simple and largely applicable.

In this paper, we focus an ensemble learning approach for rule extraction from the SVM, which uses RF rule induction technique to develop an inexpensive and possible assessment rules for diagnosis of diabetes. In our proposed method, support vectors (SVs) are primary extracted from the SVM with adequate accuracy. Then, new labels of SVs are predicted by the trained SVMmodel, and unique labels of SVs are replaced by predicted labels. At last, the fake data are fed to RF to generate rules. For extracted rule sets, if the decision tree is large, then each leaf of the tree may have little examples. On the other hand,if the tree is too small, then tree may find out few patterns. All these drawbacks make single decision tree (C4.5) difficult to in shape complex models. By utilizing the ensemble learning method, RF can answer the problem mentioned previously. Meanwhile, In view of the rule sets are generated from the SVs, the rule sets obtained by SVM + RF are definitely much less and smaller than those of RF, where the large rule sets may create the problem unintelligible. Moreover, for the skewed classification trouble the proposed method can be a preprocessing technique to decrease the imbalance proportion of skewed data, which can develop precision and recall in positive class. The model can measure undiagnosed individuals in an clear form and give a more comprehensive and obvious representation for end users.

1.1 The significance of Rule-Extraction Algorithms

The capability of representative AI systems to present a declarative demonstration of knowledge about the complexity domain offers a natural motivation capability for the decisions made by the system. Reference [3] argues that even limited explanation can absolutely influence the system's reception by

Chitra Arjun, Sri Krishna College Of Technology, Coimbatore, India
Assistant Professor Mr.Anto S, Department Of Computer Science And Engineering, Sri Krishna College Of Technology, Coimbatore, India

the user. This capability is important, mainly in the case of medical applications. A motivation capability can also offer a check on the interior logic of the system as well as being able to give a learner nearby into the problem [4]. In addition, the explanations given by rule-extraction algorithms extensively enhance the capabilities of AI systems to discover data and support the initiation and construction of new theories ANN's & SVMs have no such declarative knowledge structures, and hence, are limited in providing explanations.

1.2 The Classification of Rule-Extraction Algorithms

One possible method for classifying rule-extraction algorithms is in terms of the "translucency" of the sight taken within the rule-extraction method of the fundamental classifier.This pattern yields two crucial categories of rule-extraction techniques: " translucent " and "instructive".The distinctive feature of the "translucent" approach is that the focal point is on extracting rules at the level of entity components of the fundamental machine learning method.But in the feedforward neural networks, these are hidden and output units. Such methods obviously are used in combination with a learning algorithm that consist of rule-based explanations and the basic pattern is to use the trained classifier for generating examples for a second learning algorithm that generates rules as output [5],[6],[7]. This is the "hybrid" or "eclectic" group [1], [2], [8].Clearly, this classification scheme, initially developed for rule-extraction from neural networks, is appropriate to support vector machines as fit. Decompositional system can be based on the investigation of support vectors generated by the SVM even as learning-based classification learn come again? the SVM has learned. An example for learning-based rule-extraction from SVMs is [10].

II. Related Works

An amount of methods have been proposed for rule extraction from SVMs. Broadly speaking, these methods can be regarded as into three major families—learning based, decompositional, and eclectic method—as recommended by Andrews et al. [2] for ANNs. Learning-based method ensures the model (classifier) as a black box describing only the relationship between the inputs and the outputs. In general, learning-based approaches use another machine learning technique, which has an account capability, to study what the classifier has learned. Not like learning-based, decompositional approaches open the model, glance into its individual components, and then try to extract rules at the level of these components. Therefore, in principle, this is the most obvious approach. The eclectic approach slander in between the learning-based and decompositional approaches. The following sections review these methods.

2.1 View of Decompositional Rule-Extraction from SVMs

Reference [14] introduces an approach for rule-extraction from SVMs: the SVM+ prototype method. The fundamental initiative of this technique is to use the output decision function from an SVM and then use K-means clustering algorithm to decide prototype vectors for each class. These vectors are collective with support vectors to describe an ellipsoid in the input space which are then mapped to if-then policy. This approach does not extent well: in case of a huge number of patterns and an overlie between dissimilar attributes, the explanation ability suffers.

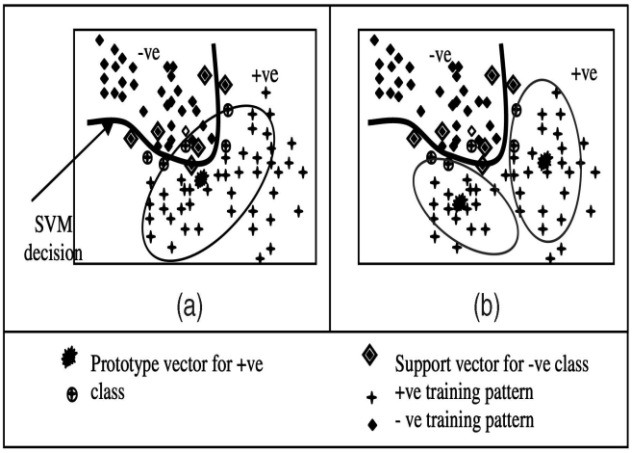

Fig1.SVM+ phases (adapted from [3]). (a) One rejoin (cluster) including outliers. (b) The n iteration after division to exclude outliers.

2.2 View of Learning-based Rule-Extraction from SVMs

References [15], [16] propose a learning-based approach for extracting rules from SVMs using two dissimilar data sets: 1) A labelled data set is used to SVM learn purposes, i.e. to construct a model with suitable accuracy. 2) A second data set is generated with the same attributes but dissimilar values to discover the simplification performance of the SVM. That is, the SVM is used to obtain the class labels for this data set. Hence a artificial data set is obtained. 3) The artificial set is then used to train a machine learning method with rationalization ability. Thereby, rules are generated that correspond to the simplification performance of the SVM.

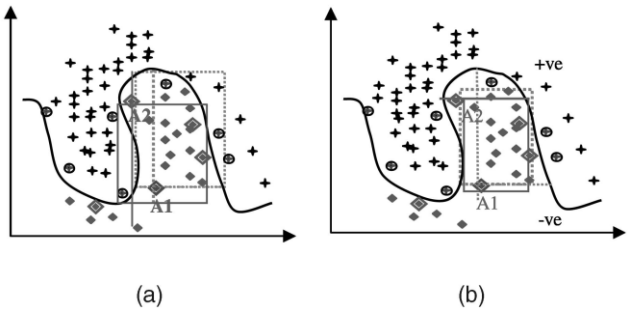

Fig.2. RulExSVM phases in a two-dimensional space (adapted from [5]). (a) Rule generation phase. (b) Tuning phase: excluding outliers.

2.3 View of Rationale behind SVM Rule Extraction

SVM rule extraction is a normal alternative of the fine researched ANN rule extraction domain. To realize the usefulness of SVM rule extraction, we need to argue 1) why rule extraction takes place and 2) why SVM rule extraction takes place rather than the further researched ANN rule extraction.

Why rule extraction,

Rule extraction takes place for the following two situations: a) to recognize the classifications made by the elementary

nonlinear black-box model, thus to "open up the black box" and b) to obtain superior performance of rule introduction techniques by removing noise in the data.

1. The most usual motivation for by means of rule extraction is to find a set of rules that can describe the original black-box model

2. An attractive study is that the better performing nonlinear model can be used in a preprocessing step to dirt-free up the data [11], [12]. By altering the class labels of the data through the class label of the black box, all outliers, that is class overlie in the data, is isolated from the data.The SVM is, as the ANN, a nonlinear predictive data mining system. Benchmarking studies have revealed that such models demonstrate very well and analogous simplification behavior [1], [13]. Still, SVMs have some significant profit over ANNs. First of all, ANNs suffer from restricted minima in the weight solution space [15].Second, several architectural choices need to be determined.Extracting rules from this state-of-the-art classification technique is the next natural step.

III. Proposed Approach

In this study, we proposed an ensemble learning approach (SVM + RF) for rule extraction from SVMs. The method applied the information provided by SVs of the SVM model, and combined ensemble techniques to extract more rules from the complex SVM model. First, C4.5, Na¨ıve Bayes Tree (NBTree), RF, and BP Neural Network were regarded as comparison methods to compare the accuracy with the SVM, which would prove the motivation of rule extraction from the SVM. Then,SVM + C4.5, an eclectic method for rule extraction, was applied to compare the rule sets' learning ability with the proposed method.SVM+C4.5 utilized the C4.5 decision tree to construct rule classifier with the same SVs. The difference between the proposed method and SVM + C4.5 was the difference in the rule induction approach. Finally, all algorithms were tested on the test sets.

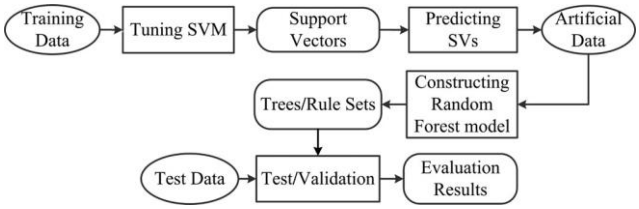

Fig.3. Block diagram: The proposed rule-extraction approach.

3.1 Data collection

In this module, the dataset is used for rule extraction and testing process. In rule-extraction process, tenfold cross validation (CV) is used as the training method to obtain the optimal parameters of models, and tenfold results integrate collectively to compute the averaged accuracy of tenfold CV for the model. The dataset contains patient's name, age and other medical report attributes. It contains 768 tuples and 8 attributes in the dataset which is used to classify the accurate results using suggested techniques. After excluding the information of individual daily food consumption, there are 56 features remained, which contained both noninvasive factors and metabolic factors. Considering the efficiency of diabetes diagnosis and screening, we proposed a detection model only with few strong relevant and easy available features. The effective feature selection and classification methods are applied to execute the efficient prediction.

3.2 Preprocessing

It is the initial process of our scenario and preprocessing is performing for improving the final dataset results more significantly. Preprocessing is the process of cleaning the database into correct format. In the given dataset, the attributes and tuples are given along with values. It analyses the data information and suggests the most appropriate transformations, missing values, replicate handling, flat pattern filtering and pattern standardization. It is used to increase the detection accuracy while classification of diabetes results. The main objective of the preprocessing step is reducing the size of volume data through filtering the irrelevant data in a specified dataset. Thus, the dataset holds unique values as well as attributes which is sued to improve the prediction performance. Also it is used to eliminate the noise data and handling missing values importantly.

3.3 Feature selection

As many machine-learning methods have worse performance with large amounts of irrelevant features, feature selection (FS) techniques have become a necessity in all applications. FS can avoid over fitting and gain a deeper insight into the unknown areas, such as occurrence and diagnosis of diseases. As a result, we utilized three filter techniques (univariate LR, chi-square tests, and information gain (IG)-based method) and an embedded technique (RF) to select the strong relevant features. Univariate LR selected the features which were statistical significant with P value < 0.05.

In statistics, chi-square test was applied to test the independence of two events. However, in FS procedure, two events represented the occurrence of the feature t and occurrence of the class ci .

$$X^2 = \frac{N[P(t,c_i)P(\bar{t},\bar{c_i}) - P(t,\bar{c_i})P(\bar{t},c_i)]^2}{P(t)P(\bar{t})P(\bar{c_i})P(c_i)} \qquad (1)$$

where N is the total number of examples in the data. (t, c_i) is the presence of t and category in c_i, ($\bar{t}$, $\bar{c_i}$) is absence of t and category not in c_i.

IG measured the information obtained for class prediction by knowing the value of a feature; the IG is defined to be the expected reduction in entropy. If features are continuous, IG uses information theoretic binning to discretize the continuous features.

The measure of feature importance in RF is the total decrease in node impurities from splitting on the variable, averaged over all trees. The node impurity is measured by the Gini importance. Gini importance is defined as

$$G_k = 2p(1-p) \qquad (2)$$

Where p represents the fraction of positive examples assigned to a confident node k and $1 - p$ as the fraction of negative examples.

3.4 Rule extraction from SVM

In this module, the unbalanced dataset is handled and data is used for training SVMs with RBF kernel. SVM is based on the principle of structural hazard minimization and it belongs to the supervised learning models for nonlinear classification analysis. The SVM model is achieved by finding the optimal separating hyperplane (w.x+b=0) with maximizing the margin *d*, which is defined as d=2/||w||. This optimal hyperplane can be represented as a convex optimization problem:

Minimize ½ $||w||^2$ subject to y_i (wx_i + b) ≥ 1 (1)

Minimize ½ $||w||^2$ + C $\sum_{i=1}^{n} \varepsilon_i$ (2)

In the nonlinear classification problem, the SVM uses kernel functions to map the examples into the high-dimensional feature space and differentiate categories by a clear linear margin. Usually, radial basis function (RBF) is used as the kernel function to map the data

$K(x, x') = \exp\left(\frac{||x-x'||^2}{2\sigma^2}\right)$ (3)

Where $||x - x'||^2$ the squared Euclidean distance between two is vectors and σ^2 is a free parameter. The kennel function becomes Hence, solving for α by the slope straight algorithm, the SVs can be obtained by the examples of training data which have nonzero Lagrange multiplier. The hyperplane is completely defined by SVs as SVs are the only examples that make contribution to the classification of the SVM. Then, the SVM model in the CV was constructed by the best fold, which was defined as the fold gave the best classification rate with the particular fold's test set, and finally the SVM model was used to test on the remained 10% dataset. To ensure the fair performance of the trained model, another nine runs were conducted on remained nine shuffled datasets with the same chosen parameters. Because on any particular randomly drawn test dataset, one classifier may outperform in testing dataset than in tenfold CV. This is a particularly pressing problem for small test datasets.

As well, if the approaches were applied to the datasets on which rule instruction system carry out better than SVM, the rule extraction from SVM would seem illogical. This aspect was always neglected in this field. In order to illustrate the motivation of rule extraction from SVM, BP neural network (BP NN), RF, C4.5, and NBTree were also implemented in ten runs as the same as SVM, whose optimal parameters were chosen by grid search in first run. The average accuracy of these models was calculated with precision, recall, F score, and AUC.

3.5 Rule generation and evaluation

The RF is an ensemble learning method for classification. RF constructs a multitude of decision trees and utilizes the mode of individual trees' output to classify the patterns. In the traditional decision tree method, it will be difficult to fit complex models (such as SVMs) if the tree is so large that each only has few examples. Unlike the decision tree, however, RF combines random subspace method and bagging idea to optimize the nonlinear problem, and it is trained based on ensemble learning, which uses multiple models to obtain better performance than any constituent model. In other words, ensemble learning, such as bagging method, can produce a strong learner which has more flexibility and complexity than single model, for instance, decision tree. Meanwhile, some ensemble methods, especially bagging, tend to reduce overfitting problems of training data, which also may intensify the generalization of the models. Totally, we utilize RF rather than decision tree to generate rule sets.

The rule generation stage proceeds in two steps: In first step, the SVM model, which is constructed by best fold of CV, is applied to predict the labels of SVs, and the original labels of SVs are discarded. Hence, the artificial synthetic data are generated. During second step, the artificial data are used to train an RF model, and all decision trees of RF are the generated rule sets. Finally, the performance of the rule sets are evaluated on 10% remained test data, the precision, recall, and F-measure are used to estimate the accuracy of the rule sets.

3.6. Performance evaluation

In this part, we contrast the existing and proposed methodologies by using c4.5, random forest, NB tree and svm algorithm. The existing system is shown the lesser performance in terms of precision, recall and accuracy values. In the proposed system, the SVM algorithm is shown the superior performance in terms of high precision, recall and accuracy values. From the experimental result, we can conclude that the proposed system is better than the existing system.

IV. Conclusion

In this section, the conclusion decides that the proposed system is provided potential performance rather than the existing scenario. In this scenario, to analyze and evaluate the Diagnosis of Diabetes we introduced an ensemble approach. The method named as support vector machine with random forest which is rule extraction based concept. Identifying the potential individuals with undiagnosed diabetes is the basic intention of this scenario. Support vectors to predict the labels by obtained SVM model, and overrode the original labels of SVs to create synthetic artificial data. Then, artificial data generated by SVM were antinormalized, and used to construct rule sets based on RF, whose process is achieved by R with package random forest. The optimal parameters of ntree (number of trees to grow), mtry (number of variables at random sampled as candidates at each divide), and node size (minimum size of terminal nodes in each tree) in RFs. The proposed result shows that our proposed model has high quality in terms of diagnosis with precision, which meant the diagnosis ability of the model. The experimental result concludes that the proposed system is superior to existing system

References

[1] A.B. Tickle, R.Andrews, M.Golea, and J.Diederich, "The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural network", IEEE Trans. Neural Networks, vol. 9(6), pp. 1057-1068, 1998.

[2] R. Andrews, J. Diederich, and A.B. Tickle, "A Survey and Critique of Techniques For Extracting Rules From Trained Artificial Neural Networks", Knowledge Based Systems, vol. 8, pp. 373-389, 1995.
[3] R. Davis, B.G. Buchanan, and E. Shortcliff, "Production Rules as a Representation for a Knowledge Based Consultation Progra", J. Artificial Intelligence, vol. 8(1), pp.15-45, 1977.
[4] S. Gallant, "Connectionist Expert System", Communications of the ACM, vol. 31 (2), pp. 152-169, 1988.
[5] S. Sestito and T. Dillon, "Automated Knowledge Acquisition of Rules With Continuously Valued Attributes", in Proc.12th International Conference on Expert Systems and their Applications (AVIGNON'92), Avignon -France, 1992, pp. 645-656.
[6] M.W. Craven, and J.W. Shavlik, "Using Sampling and Queries to Extract Rules From Trained Neural Networks", in Proc. of the 11th International Conference on Machine learning, NJ, 1994, pp.37 45.
[7] G. Towell, and J. Shavlik. "The Extraction of Refined Rules From Knowledge Based Neural Networks", J. Machine Learning, vol. 131, pp.71-101, 1993.
[8] M.W. Craven, and J.W. Shavlik, "Extracting Tree–Structured Representation of Trained Networks", Advances in Neural Information Processing Systems, vol. 8, pp.24-30, 1996.
[9] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.Cambridge Univ. Press, 2000.
[10] U. Johansson, R. Ko¨nig, and L. Niklasson, "The Truth Is in There—Rule Extraction from Opaque Models Using Genetic Programming," Proc. 17th Int'l Florida AI Research Symp. Conf. (FLAIRS), 2004.
[11] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines," European J. Operational Research, vol. 183, no. 3, pp. 1466-1476, 2007.
[12] B.D. Ripley, "Neural Networks and Related Methods for Classification," J. Royal Statistical Soc. B, vol. 56, pp. 409-456, 1994.
[13] J. Huysmans, B. Baesens, and J. Vanthienen, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models," K.U.Leuven KBI, Research 0612, 2006.
[14] T. Van Gestel, J. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle, "Benchmarking Least Squares Support Vector Machine Classifiers," Machine Learning, vol. 54, no. 1, pp. 5-32, 2004.
[15] C. Bishop, Neural Networks for Pattern Recognition. Oxford Univ. Press, 1996.
[16] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, Least Squares Support Vector Machines. World Scientific, 2002